GPT-3 and InstructGPT: technological dystopianism, utopianism, and “Contextual” perspectives in AI ethics and industry

Year: 2022 | Citations: 110 | Authors: Anastasia Chan

Abstract

This paper examines the ethical solutions raised in response to OpenAI’s language model Generative Pre-trained Transformer-3 (GPT-3) a year and a half from its release. I argue that hype and fear about GPT-3, even within the Natural Language Processing (NLP) industry and AI ethics, have often been underpinned by technologically deterministic perspectives. These perspectives emphasise the autonomy of the language model rather than the autonomy of human actors in AI systems. I highlight the existence of deterministic perspectives in the current AI discourse (which range from technological utopianism to dystopianism), with a specific focus on the two issues of: (1) GPT-3’s potential intentional misuse for manipulation and (2) unintentional harm caused by bias. In response, I find that a contextual approach to GPT-3, which is centred upon wider ecologies of societal harm and benefit, human autonomy, and human values, illuminates practical solutions to concerns about manipulation and bias. Additionally, although OpenAI’s newest 2022 language model InstructGPT represents a small step in reducing toxic language and aligning GPT-3 with user intent, it does not provide any compelling solutions to manipulation or bias. Therefore, I argue that solutions to address these issues must focus on organisational settings as a precondition for ethical decision-making in AI, and high-quality curated datasets as a precondition for less harmful language model outputs.